|  |  |  |
| --- | --- | --- |
| http://www.abdn.ac.uk/iahs/slideshow-images/DaSH_logolong2_rdax_356x267.jpg  **Grampian Data Safe Haven (DaSH)** | | |
| ***Developing a Data Specification*** | Version: 1.1  Date: 13/04/15 | Approved By:  Date: DD/MM/YY |
| Owner:  *Steph Hall* | Next Review: |

Document History and Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Description** |
| 1.00 | 13/04/2015 | Steph Hall | Document creation & inclusion of feedback from CB |
|  |  |  |  |

# Developing a Data Specification

Developing a data specification is an important part of any research that requires use of an extract of data from a larger dataset. It provides a means of communicating clearly to data custodians and data linkage analysts exactly what data variables are required and how the file of extracted data should be structured (eg one person’s records in one row or multiple rows).

The development process may be iterative and involve discussions with the data custodians, the safe haven data analyst and a Research Co-ordinator as well as your research team. Use version control on your documents to avoid confusion.

## Frequently asked questions

1. **How do I find out what datasets are available?**

A good starting place for national data is the Administrative Data Liaison Service (ADLS) for social, economic and health data ([ADLS](http://www.adls.ac.uk/find-administrative-data/)) or ISD Scotland’s (ISD) National Data Catalogue ([NDC](http://www.ndc.scot.nhs.uk/National-Datasets/index.asp)) for NHS Scotland’s health and health related data. For datasets local to Grampian see the University of Aberdeen’s web site ([UoA datasets](http://www.abdn.ac.uk/iahs/research/health-informatics/datasets.php)) or contact a member of the Health and Data Linkage in North East Scotland ([HEADLINES](http://www.abdn.ac.uk/iahs/research/membership-225.php)) network.

1. **How do I find out what data is available in a dataset of interest?**

Online data dictionaries for the datasets listed by ADLS are available on the *Administrative data information packs* web page ([Info packs](http://www.adls.ac.uk/information-packs/)), this includes ISD data dictionaries. Alternatively for ISD data advice contact their Electronic Data Research and Innovation Service ([eDRIS](http://www.isdscotland.org/Products-and-Services/eDRIS/)). For local datasets contact the Data Custodians for assistance; contact details can be found on the web ([UoA datasets](http://www.abdn.ac.uk/iahs/research/health-informatics/datasets.php)). A DaSH Research Co-ordinator will also be able to recommend who to contact.

1. **Does it matter if I request more or less data than I think I need?**

Yes, you need to be specific. Request only the data you require to answer your research questions. If you find you need additional variables after your project has received all the relevant permissions (eg ethics, data access approval) you will be required to submit an amendment to your application to the permission providers which will take time to process.

Some datasets are very large so you could end up with an extract that is difficult to understand and deal with. Additional work would be required to make it manageable and appropriate to your project and the computing power you have available. Good data specification is a key part of planning your project

1. **How do I know which variables to use for linking the different datasets?**

Linking datasets using a unique identifier for each person/event is the simplest form of linkage (this is a type of deterministic linkage a rule-based exact one-to-one matching of characters in the linkage variable(s)). NHS Scotland’s Community Health Index (CHI) numbers are commonly used for linking records across health datasets. In the deterministic method the importance of the linkage variables is usually equal whereas in probabilistic linkage the variables differ in their quality and discriminatory power. An algorithm is used in the latter to calculate the probability that records belong to the same person. A combination of linkage variables eg *surname-date of birth-postal code* is called the linkage key.

There are pros and cons to both methods and the choice will depend on which variables are available and the quality of those variables. Seeking advice from an appropriate person (eg DaSH analyst) is strongly recommended.

1. **I need to generate derived variables from the raw data in my data extract for my analysis can I get help with this?**

Yes. A DaSH analyst can help you by creating new variables if your requirements are clearly defined.

1. **Do I need to provide codes for diseases and drugs that will be in my data extract?**

Yes. You can provide either Read or ICD-9 or ICD-10 code lists for the diseases in your study. Drugs can be specified using generic or brand names or British National Formulary codes. Speak to a member of the DaSH team for advice.

1. **Can I get advice on how my data extract file should be structured?**

Yes. This can be discussed with a DaSH analyst. The options are one row of data per patient or event/observation (wide format) or multiple rows (long format). Your choice will depend on what your unit of analysis is and what analyses you will be undertaking. Data may also be held in one table or multiple tables.

## eDRIS

The national safe haven, eDRIS is able to offer advice for projects using national health and health related data. For more information see [eDRIS](http://www.isdscotland.org/Products-and-Services/eDRIS/).

## Data specification

Useful things to include:

1. names of the source datasets your data will be extracted from
2. inclusion and exclusion criteria that define the subset or cohort of patients to be extracted from each source dataset
3. dates on which inclusion criteria have to be met
4. time period the data extract is to cover
5. any rules on how you want the data to be handled or cleaned
6. names you want assigned to the variables
7. how each variable is defined or calculated especially if they are derived variables
8. how variables should be coded
9. data type of each variable
10. disease codes (Read, Quality and Outcomes Framework (QOF), International Classification of Diseases (ICD-9 or ICD-10))
11. drug names (generic &/or brand) and/or codes (British National Formulary (BNF)).
12. file layout

## Examples data specifications

Three examples of data specifications are provided to illustrate the key features of a data specification listed above.

Example one**: Medication adherence study (non-linkage)**

***Overview***

Anonymised primary care data was extracted from the Primary Care Clinical Informatics Unit Research (PCCIUR) dataset to create a cohort of diabetes patients, one prescribed ACE inhibitors. Data on patient demographics, clinical diagnoses and comorbidities was extracted along with patient medication histories (dates and durations of prescriptions) for the study drugs.

|  |  |
| --- | --- |
| **Project title**  **[DaSH reference]** | **Adherence to medication in diabetes patients**  [n/a] |
| **Research team** |  |
| **Source dataset(s), location** | PCCIUR, University of Aberdeen |
| **Study period** | 1st April 2005 up until the latest research database update in 2010 |
| **Inclusion criteria** | **Criteria to be met on 1st April 2005:** patients must   * have a diagnostic code for diabetes mellitus (DM) consistent with the QOF register *[list of disease codes(Read) provided for the linkage analysts]* * be 17 years of age or over * be alive and fully registered with a general practice   **Prescription related criteria**   * Patients must have at least one prescription for a drug of interest ie an ACEi during the study period (ie on or after the 1st Apr 2005) *[list of BNF codes/drug names provided for the linkage analysts]*. If a patient meets the above inclusion criteria but has switched drugs at some point they can still be included as long as the replacement drug is in the same class as the original drug of interest * Patients must be fully registered with the practice for one year prior to the ‘index prescription date’ (see definition below) |
| **Exclusion criteria** | n/a |
| **Definitions** | **Index prescription date**  The date of the first prescription for a medication of interest which falls on or after the 1st of April 2005 (a person may have scripts for the medication of interest prior to the start of the study period).  **Index end data**  Index prescription date + 364 days. Period during which medication adherence will be assessed. |
| **Data handling rules** | 1. If one or more prescriptions for the same drug is/are issued on the same date or over 2 consecutive days treat them as one script and sum the quantity dispensed to get the total. |
|  | 1. Allow 15 columns for the prescription variables. Every patient will have index prescription variables plus variables for the 14 refills (even if some have no data in them). |
| **File layout** | **ACEI cohort**: flat file containing one record per patient |

***Variables, disease coding and drug list***

**Table 1: ACEI cohort – demographics, diagnoses, index readings and comorbidities** (a sample of variables only)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1a. Demographics** | | | | |
| **Full field name** | **Data type** | | **Description** | **Coding** |
| pt\_ID | integer | | unique identifier for each patient | n/a |
| sex | integer | |  | 1=male; 2=female |
| dob | date | | date of birth | dd/mm/yyyy |
| death\_date | date | | date of death | dd/mm/yyyy |
| SIMD | integer | | deprivation quintile | n/a |
| **1b. Diagnoses** | | | | |
| diabetes\_date | date | | date of first code for diabetes (Read or QOF) | dd/mm/yyyy |
| diabetes\_type | integer | | type of diabetes at date of first code | 1=type 1; 2=type 2; 3=not recorded |
| **1c. Index readings** | | | | |
| syst | integer | any systolic blood pressure readings at any time; this is anytime in PCCIU not just within the study period | | 0=no; 1=yes |
| syst\_date\_index | date | date of the systolic reading nearest to and **prior** to ACEi index date | | dd/mm/yyyy |
| syst\_index | integer | systolic blood pressure reading nearest to and **prior** to ACEi index date | | n/a |
| smoking\_date\_index | date | date of record of smoking status nearest to (and prior to) ACEi index date | | dd/mm/yyyy |
| smoking\_status\_index | integer | smoking status nearest to (and prior to) ACEi index date | | 0= non smoker; 1= smoker; 2=ex-smoker;  3=no record |
| HbA\_date\_index | date | date of HbA1c reading nearest to (and prior to) ACEi index date | | dd/mm/yyyy |
| HbA\_index | integer | HbA1c reading nearest to (and prior to) ACEi index date | | Value/missing value |

|  |  |  |  |
| --- | --- | --- | --- |
| **1d. Comorbidities** | | | |
| CVD | integer | any code for coronary vascular disease (CVD) at anytime | 0=no; 1=yes |
| CVD\_date | date | date of first code (Read or QoF) for CVD | dd/mm/yyyy |
| CKD | integer | any code for chronic kidney disease (CKD) | 0=no; 1=yes |
| CKD\_date | date | date of first code (Read or QoF) for CKD | dd/mm/yyyy |
| **1e. Polypharmacy** | | | |
| polypharm\_index | integer | count of all items prescribed in a 3 month period prior to the index\_date; duplicate prescriptions on the same day or on two consecutive days should be counted as one item | na |

**Table 2: ACEI medication histories – index prescription and first to 14th refill**

|  |  |  |  |
| --- | --- | --- | --- |
| **INDEX PRESCRIPTION** |  |  |  |
| **Full field name** | **Data type** | **Description** | **Coding** |
| index\_date | date | Date of 1st prescription of an ACEi on or after 1st April 2005 with no prescriptions for the same or any other ACEi in preceding 6 months. | dd/mm/yyyy |
| index\_drug | string | Name of the ACEi prescribed on the index\_date | n/a |
| index\_enddate | date | index\_date + 364 days | dd/mm/yyyy |
| index\_PD | integer | Prescription duration (days) of index drug = quantity of tablets in prescription / ACE\_DDC. Where ACE\_DDC is the daily dose count (ie number of tabs per day = # tabs/dose X daily dosing frequency). | na |
| index\_days2next | integer | Length of time (days) between the index script date and the actual date of Refill1 (i.e. refill1\_act\_date: see table below). Refill1\_act\_date – Index\_date | na |

|  |  |  |  |
| --- | --- | --- | --- |
| **FIRST REFILL ie 1st script after index script** |  |  |  |
| **Full field name** | **Data type** | **Description** | **Coding** |
| refill1\_act\_date | date | Actual date of 1st refill | dd/mm/yyyy |
| refill1\_PD | integer | Prescription duration (days) of first refill = quantity of tablets in refill1 script / ACE\_DDC. Where ACE\_DDC is the daily dose count (ie number of tabs per day = # tabs/dose X daily dosing frequency). | na |
| refill1\_ex\_date | date | Expected date of 1st refill = index\_date + index\_PD | dd/mm/yyyy |
| refill1\_days2next | integer | Length of time (days) between the actual date of refill2 and the actual date of refill1 (refill2\_act\_date – refill1\_act\_date) | na |
| refill1\_delta | integer | The difference in days between refill1\_ex\_date and refill1\_act\_date | na |
| **Repeat the above variables up to and including refill number 14.** | | | |

**Table 3: Rules**

|  |
| --- |
| 1. If one or more script for the same drug is issued on the same date or over 2 consecutive days treat them as one script and sum the quantity dispensed to get the total. 2. Every patient will have index variables then variables for the 14 refills even if some have no data in them |

**Table 4: BNF drug codes and generic names of drugs to be included in the study (part of table shown)**

|  |  |  |
| --- | --- | --- |
| **British National Formulary (BNF) 60 Sept 2010** | | |
| **Drug class 2.5.5.1**  **ACE inhibitors** | | **Drug class 2.12**  **statins** |
| **captopril** | *Lisicostad 20/12.5* | **atorvastatin** |
| *Captopril* | *Zestoretic* | *Lipitor* |
| *Ecopace* | *Zestoretic 10* | **fluvastatin** |
| *Kaplon* | *Zestroretic 20* | Fluvastatin |
| *Tensopril* | **moexipril hydrochloride** | *Lescol* |
| *Capoten* | *Perdix* | *Lescol XL* |
| *Co-zidocapt* | **perindopril erbumine** | **pravastatin sodium** |
| *Capto-co* | *Perindopril* | *Pravastatin* |
|  |  |  |

**Table 5: Read codes for diabetes: (DISEASE CODE OR DRUG CODE)**

14F4.00 - H/O: ADMISSION IN LAST YEAR FOR DIABETES FOOT PROBLEM

14P3.00 - H/O: INSULIN THERAPY

2BBF.00 - RETINAL ABNORMALITY - DIABETES RELATED

2G51000 - FOOT ABNORMALITY - DIABETES RELATED

2G5A.00 - O/E - RIGHT DIABETIC FOOT AT RISK

2G5B.00 - O/E - LEFT DIABETIC FOOT AT RISK

66A3.00 - DIABETIC ON DIET ONLY

66A4.00 - DIABETIC ON ORAL TREATMENT

66A5.00 - DIABETIC ON INSULIN

66AA.11 - INJECTION SITES - DIABETIC

66AH000 - CONVERSION TO INSULIN

66AI.00 - DIABETIC - GOOD CONTROL

66AJ.00 - DIABETIC - POOR CONTROL ………………

Example 2: **Data specification for anonymised data about pregnant mothers held in the Aberdeen Maternal and Neonatal Database**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Full Field name** | **English description** | **Data Type** | **ICD-9** | **Coding of variable** |
| **Age** | Age at the time of delivery | Integer | n/a |  |
| **Location** | Whether the mother is resident within Aberdeen City | Boolean | n/a | 1=yes  0=no |
| **SC\_Husb** | Social class of husband | Integer | n/a |  |
| **SC\_PM** | Social class pre-marital | Integer | n/a |  |
| **Dep\_Cat** | Deprivation category | Integer | n/a |  |
| **Fath\_ID** | For each delivery a sequential ID indicating whether the DOB of the father is the same or a different DOB | Integer | n/a |  |
| **Ethnic** | Ethnicity as registered | String | n/a |  |
| **Smok** | Whether a smoker at the time of pregnancy | Boolean | n/a | 1=yes  0=no |
| **Height** | Height of mother (in cm) | Integer | n/a |  |
| **Weight** | Weight of mother (in kg) | Integer | n/a |  |
| **Wei\_Date** | At what week of gestation was the weight of the mother recorded | Integer | n/a |  |
| **Para** | What is the mothers maximum parity recorded in the database | String | n/a |  |
| **Num\_Preg** | What number of pregnancies does the mother have recorded in the database | Integer | n/a |  |

Example 3: **Data specification for anonymised data relating to infertility**

| **VARIABLE** | | **FULL NAME** | **DESCRIPTION** |
| --- | --- | --- | --- |
| **RowID** | | Row identifier | An identifier for the date ordered rows per woman |
| **Date** | | Date of event | Date of event in each row e.g. date of first visit, date of treatment cycle etc. in ddmmyy10. format  NOTE: For ARUD use date of embryo transfer as ‘Date’  NOTE: Is cause of infertility diagnosed at first visit or does it have a separate date? |
| **Baseline variables from AFCD:** | |  | Please repeat over each line per woman |
|  | Fage | Female age | In years please |
|  | Mage | Male age | In years please |
|  | DurInf | Duration of infertility | In months if possible |
|  | Finf | Female infertility | Code as: Primary=1; Secondary=0 |
|  | Minf | Male infertility | Code as: Primary=1; Secondary=0 |
|  | RefClinic | Referral clinic |  |
|  | RefBy | Referred by GP/Hospital | GP=1; Hospital=0 |
|  | Amen | Amenorrhoea | Code as: Primary=1; Secondary=0 |
|  | Region | Region of referral | Code as: Grampian/Highland/Orkney=1; Shetland/Other=0 |
|  | FLastFert | Time since last female fertilisation | In months if possible |
|  | MLastFert | Time since last male fertilisation | In months if possible |